Mathematics Subject Classification and related schemes in the OAI framework

Antonella De Robbio, Dario Maguolo
Mathematics Library – University Library System University of Padova – ITALY
Alberto Marini
Institute for Applied Mathematics and Information Technology – National Research Council (CNR-IMATI), Milano - ITALY

Electronic Information and Communication in Mathematics
Beijing, August 29-31, 2002
A satellite conference to the ICM 2002, International Congress of Mathematicians

1. Introduction

This talk aims to give a feeling of the roles that discipline-oriented subject classifications play in scientific communication, in the perspective of the Open Archives movement for the free dissemination of information in research activities.
Mathematics, and Mathematics Subject Classification, will be the focuses around which we will move to discover a variety of presentation modes, protocols and tools for human and machine interoperability.
The Open Archives Initiative (OAI) is intended to be the effective framework for this play, which is presented to you by two university librarians at Padova and a researcher at Milano, Italy.

2. Contents

The talk is divided into two parts.
In the first one, we start by giving some examples of the most important subject classification schemes in mathematics and related disciplines. Then we sketch the structure of subject classification schemes in view of browsing. Finally we give an idea of different browsing modalities, which are demonstrated by the tools we produced and collected in *The Scientific Classifications Page*.
In the second part, we give an account of different strategies for e-print communication in scientific research, up to the basic definitions of the Open Archives Initiative.
A review of the functionalities actually implemented in OAI compatible archives managed by the EPrints software will lead us to some working hypothesis about the roles that subject classifications in mathematics and related disciplines can play in the scenarios of the Open Archives movement.

3. Subject classification schemes in mathematics

Subject classification schemes are primary tools for the organization of knowledge and terminology in scientific disciplines. Mainly they are produced by professional societies, or academic and research institutions, for use in their own bibliographic databases.
Mathematics Subject Classification covers all branches of pure and applied mathematics, including probability and statistics, numerical analysis and computing, mathematical physics and economics, systems theory and control, information and communication theory.

On the side of mathematics education, we have the **Zentralblatt für Didaktik der Mathematik Classification Scheme**.

4. Subject classification schemes in computing and physics

In the field of computing, the most important tool is the **ACM Computing Classification System**. Section 68 *Computer Science* of MSC was designed in rather tight matching with a great part of CCS.
In the fields of theoretical, experimental and applied physics and astronomy we have the **Physics and Astronomy Classification Scheme**. Section 02 *Mathematical methods in physics* of PACS closely resembles the top level codes for pure mathematics, probability and statistics of MSC.
A version of **PACS** is established as **Section A** of **INSPEC Classification**.
The **fields of economics** are increasingly involved in mathematical arguments, both in theoretical and specific topics; and conversely, mathematical problems and theories even more often arise from economic domains. See **Journal of Economic Literature Classification System**, developed by the American Economics Association for its indexing journal and for the corresponding *EconLit* database.
Such topics are mostly located in the 62 *Statistics*, 90 *Operations research, mathematical programming*, and 91 *Game theory, economics, social and behavioral sciences* sections of MSC2000.

5. The common structure of subject classification schemes

Besides these, many other subject classification schemes exist for use in any scientific disicpline or field of disciplines. Yet other schemes are the general ones, not oriented to specific disciplines, such as Dewey Decimal Classification.
Anyway, the structure is essentially the same: a relational system of *categories*, identified by alphanumerical *codes*, whose meaning is specified by *descriptions* or scope notes in some natural language.
Generally there is one main relation, which in most cases is tree-shaped. Sometimes, however, the main relation is a more relaxed partial order, allowing nodes to be under more than one node (so the relation is called multihierarchical).
Other relations are considered as cross-references, allowing connections between diverging paths of the main relation.
Subject classification schemes vary in time through succeeding versions; usually one version keeps valid for indexing and searching in a bibliographic database for a more or less long period of years.

6. Classification schemes: from structure to browsing

Due to their structural features, subject classifications are effective tools for browsing and searching in bibliographic databases, catalogs and other kinds of metadata repositories.
Moreover, subject classifications can set up knowledge organization tools for lexical collections extracted from metadata or fulltext databases, for terminologies, glossaries, dictionaries or encyclopedias, surveys, up to distributed libraries of natively digital documents or digitalized paper document. The set of descriptions of a classification scheme is itself a primary terminological resource.

7. The Scientific Classification Page

Different modes in browsing subject classifications can be exploited by hypertextual techniques. We managed to produce various tools to demonstrate some of these modes.
The Scientific Classifications Page collects such tools. It is presented both in English and in Italian language. It includes the following sections:

- *The Mathematics Classification Page*
- *Mathematics Subject Classification MSC and Dewey Decimal Classification DDC*
- *Relating Scientific Subject Classifications*

and three display modes:

- Simple frame
- Double view
- KWIC lists of descriptions

8. Our tools

The tools we produced consist of systems of syntactically simple but highly connected and coordinated HTML pages, called *h-volumes*.
H-volumes can amount even to thousands of files, written in plain HTML with simple JavaScript routines; in our working environment they are generated by a pool of standard C programs, starting from ASCII files, which present lists of records without redundancies and glossaries concerning attribute values.
H-volumes can be employed to display any kind of structured information set, such as directories, biographical collections, metadata collections, databases, glossaries, dictionaries, encyclopedias.
The actual production of h-volumes starts from ASCII files obtained by manipulating existing data sets and texts, in particular available Web pages. This preparation activity is worked out partly by hand (i.e. using interactively some flexible source editor), partly making use of text processing procedures developed contextually to the development of procedures for HTML page generation.

9. The Mathematics Classification Page

Let's now turn to see the sections of ***The Scientific Classifications Page*** in some detail.
The Mathematics Classification Page collects six hypertextual frame presentations (five simple frame and one double view) of the latest version of Mathematics Subject Classification, MSC2000.The simple frame presentations display the classification with descriptions expressed in English, in Italian, with interleaved descriptions in English and Italian, with marks of changes from MSC 1991, with links to subject-specific guide pages for relevant Websites.

10. The Simple Frame Presentation

This is an example of simple frame presentation.
The top frame is a sort of Table of Contents, which gives access to different sliciings of the scheme: single list presentations of the classification categories at level 1 and 1-2, and an indexed set of list presentations which covers the whole scheme.
For the latter, the top frame displays the list of the first 2 digits of the codes of the 63 level 1 categories; each item in the list points to a page which is displayed in the frame below, containing a list presentation of the subtree below the indicated level 1 category.

11. Mathematics Subject Classification MSC and Dewey Decimal Classification DDC

On the other hand, double or multiple view presentations can be exploited to navigate through transversal links either inside one version of a classification scheme or among more schemes or versions: you can move to and from parallel views of them.
Here is an example of double view presentation, showing connections between classification categories from the Dewey Decimal Classification, 21st edition, and MSC2000.
This section includes also
• a KWIC list h-volume for the combined set of descriptions

12. KWIC KeyWords In Context list H-volumes

KWIC list h-volumes are devised for discovering textual similarities among subject descriptions in one or more classification schemes or versions, in order to obtain suggestions about possible affinities of contents.
A KWIC list presents every description through its circular permutations, beginning with a significant word or phrase; the overall list is ordered along the list of significant words.
By a method similar to that employed for simple frame presentation, long ordered list, as generally a KWIC list is, can be endowed with some sort of distributor allowing to reach quickly determined points or sections of the long sequence. A distributor can be built with pointers to initial letters, initial words of paged sections, sublists dealing with particular categories of entities. The list of permuted descriptions, subdivided into smaller manageable lists, is displayed on the right, while the distributor appears in the left frame.

13. Relating Scientific Subject Classifications

The ***Relating Scientific Subject Classifications*** section of ***The Scientific Classifications Page*** contains a set of English language presentations (in one case bilingual):

- a double view presentation, showing double view presentation, showing connections between classification categories from ACM Computing Classification System (1998) and MSC 2000
- separate KWIC lists of descriptions of MSC 2000, of PACS 2001, of ACM Computing Classification System (1998)
- combined KWIC lists of descriptions of MSC 2000 and PACS 2001, and of MSC 2000 and ACM Computing Classification System (1998).

14. The prototype

The h-volumes we produced are not intended to be taken as ultimate references, but as prototypes capable to clarify the real problems to face for the production of more complete and professional h-volumes and to test their effectiveness as documentation tools.
In fact, the development of such prototypes brought to the specification of parametrization mechanisms, data structures and processing modes which induced to define a programming

language oriented to the manipulation of hypertextual presentations and displays of mathematical structures.
On the other end, KWIC lists may not be intended for the end user, rather as a help in establishing structured connections within or among classification schemes. This activity, although can greatly benefit from automated techniques, requires an amount of field specific knowledge which can't be automated, at least with the current technologies.
The connections so discovered can be subsequently displayed through multiple view presentations of the involved classification schemes.

15. Towards the OAI framework

In the second part of this talk, we will sketch some perspectives and working hypotheses about the roles that subject classifications in mathematics and related disciplines can play in the scenarios of the Open Archives movement for the free dissemination of information in research activities.
Scientific research relies heavily on the rapid dissemination of results. So the slow formal process of submitting papers to journals has been augmented by other, more rapid, dissemination methods. Originally these involved printed documents, such as technical reports and informal conference papers...

16. Searching through personal homepages and small archives

Then researchers started taking advantage of the Internet, putting papers on ftp sites and later on various web sites. But these resources were fragmented. Searching through them resulted to be very difficult, and there was no guarantee that information would be archived at the end of a research project.
Harvest and cash engines like Citeseer, then Researchindex, provided a solution which has been appreciated especially by people in the computing area.

17. E-print communication: tools and networking strategies

Other strategies for scientific research communication via e-prints involve:

- small specialized archives
- a potentially catch-all centralized archive such as arXIv for physics and related disciplines, mathematics, nonlinear sciences, computational linguistics, and neuroscience
- single or networked institutional archives, such as NCSTRL and the ERCIM Technical Reference Digital Library for computer science and mathematics
- distributed networks connected by some interoperability protocol, such as RePEc for economics, ReLIS for library and information science
- umbrella servers, such as MPRESS for mathematics
- servers connected to groups of journals or sponsored by commercial publishers

18. The Open Archives Initiative

The Open Archives Initiative is an international effort to develop interoperability standards for disseminating content over the Web.

OAI stresses the separation of being a *data provider*, that is managing an archive, and being a *service provider*, that is, an interface for search, browsing, reference linking) On the other hand, nothing prevents the same system to embody and integrate both functions.It is even possible for individual researchers to develop personal open archives, which can be accessed to build tailored personal web sites and other services, as well as harvested into department archives.
The base concept of the OAI is metadata harvesting, which is realized in the **OAI Protocol for Metadata Harvesting**.
So it no longer matters *where* papers are archived; the papers in all registered OAI-compliant archives can be harvested using the OAI protocol into one global "virtual archive" by Open Archives service providers.

19. OAI compatible refereed self-archives: the *EPrints* software

EPrints is a free (General Public License) software for managing e-prints archives, developed at the Electronics and Computer Science Department of the University of Southampton (UK). It is aimed at organizations and communities rather than individuals. It provides an interface for system administrators, for archive editors to process submissions, for authors to deposit papers, and for users to access papers by searching or browsing metadata.
The system comes configured to run an institutional pre-prints archive, but can be reconfigured with utterly different metadata fields and content.
Any version of EPrints is fully interoperable with the current OAI Protocol for Metadata Harvesting.

20. Conclusions

Our work has been directed to the definition of text processing methodologies for the development of hypertextual presentations of complex documentation structures.
Such presentation modalities can enrich the browsing functionalities of archives and service providers in the OAI framework, allowing a full network of bridges among specific subject areas to guide advanced research communication activities.
In particular, we are investigating the possibility of providing the EPrints software with tools modeled on the experimental ones we produced for the *Scientific Classification Page*.
Centering with Mathematics Subject Classification, we can launch bridges and pass them through inside mathematics and among the disciplines that live and develop with mathematics. This amounts to have bridges launched all over the world of scientific and technological knowledge, if we are aware of the dynamics that mathematical disciplines are ever more moving in modeling and computing activities for every field of human knowledge.